Rosanoff

# A PRACTICAL SIMPLIFICATION OF THE METHOD OF LEAST SQUARES

by

M. A. Rosanoff, Sc.D.



A Lecture given at the

Galois Institute of Mathematics

at

Long Island University

300 Pearl Street, Brooklyn, N. Y.

# A PRACTICAL SIMPLIFICATION OF THE METHOD OF LEAST SQUARES.

BY M. A. ROSANOFF, Sc.D., PROFESSOR OF CHEMICAL RESEARCH
AND DEAN OF THE GRADUATE SCHOOL, DUQUESNE UNIVERSITY,
PITTSBURGH, PENNSYLVANIA. *

In course of our researches on partial vapor pressures and the theory of distillation,† my students and myself had frequent occasion to use the Method of Least Squares. To facilitate the extensive computations involved, I devised a simplification and calculated a number of auxiliary formulae, which may save much superfluous labor to others and are therefore reproduced in the following pages.

In the mathematical treatment of scientific results labor is often wasted on a degree of precision in excess of the accuracy of the results themselves. For instance, two experimental figures, 6.7893 and 3.4578, involving an error of at least 1 part in 35,000, might be multiplied with arithmetical rigor to obtain the product 23.47604154 implying an error of 1 part in two billion. An observer of experience, aiming merely to keep his mathematics within the limits of his experimental errors, would multiply the two figures in some such way as this:

$$\begin{array}{r} 6.7893 \\ \underline{3.4578} \\ 20.3679 \\ 2\ 7157 \\ 3395 \\ 475 \\ \underline{54} \\ 23.4760 \end{array}$$

In the product, written 23.476, the multiplication error of 4 in two million

*Communicated by the Author.
†Partly summarized in Sydney Young's Distillation (Macmillan & Co., London and New York, 1922.)

would be negligible compared with the experimental error of the multiplier.

In the use of the Least Squares this type of simplification must not be employed without alert watchfulness, matters being complicated by the additions and subtractions, by which the relative errors are liable to be greatly magnified. The semi-graphic procedure recommended below, vaguely analogous in that it too aims merely to keep the mathematical errors within those of the experiments to be represented, will be found accurate enough for all ordinary purposes and safe. The procedure is based on the substitution of carefully interpolated figures for the actual results of observation.

The given experimental results are plotted on accurately ruled millimeter paper, the scale large enough to show the likely errors of observation or experiment. A smooth curve is drawn free-hand to represent the trend of the points as closely as the eye will allow; or else a neat wavy curve is drawn through the points themselves. In some cases, if the points are more or less evenly thrown by the errors to the one and the other side of any curve on which they might belong, neighboring points may, for the purpose of interpolation, be connected by straight lines. From the curve, or from the broken line, we read the ordinates corresponding to a set of uniformly increasing abscissas to which are assigned the values $x = 0, 1, 2, 3, 4, \ldots 9, 10$; or some similar set. Still further simplification comes of assigning to the abiscissas the values $(x-5) = -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5$.

These ordinates yield $\sum y$, and simple further calculation leads to the values of $\sum xy, \sum x^2y$, etc. The equations that would ordinarily result have been solved by me in advance for the coefficients a, b, c, ... of a series of equations of the form

$$y = a + bx + cx^2 + \ldots$$

would be negligible compared with the experimental error of the multi-plier.

In the use of the Least Squares this type of simplification must not be employed without alert watchfulness, results being readily caused by the additions and subtractions, by which the relative errors are likely to be greatly magnified. The semi-graphic procedure recommended below, wholly analogous in that it too aims merely to keep the mathematical errors within those of the experiments to be represented, will be found accurate enough for all ordinary purposes and safe. <u>The procedure is based on the substitution of smoothly interpolated figures for the actual results of observation.</u>

The given experimental results are plotted on accurately ruled millimeter paper, the scale large enough to show the likely errors of observation or experiment. A smooth curve is drawn free-hand to represent the trend of the points as closely as the eye will allow; by which I only mean curve is drawn through the points themselves. In such cases, if the points are more or less evenly thrown by the errors to the one and the other side of any curve on which they might belong, the plotted points may, for the purpose of interpolation, be considered as standing there. From the curve, we read the tangent line, we read the ordinates corresponding to a set of uniformly increasing abscissae to which are assigned the values $x = 0, 1, 2, 3, 4, \ldots, 9, 10$; or some similar set. Still further simplification occurs if, particularly in the abscissae the values $(x - x_0) = -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5$.

These ordinates yield $y_0$, and simple further calculation leads to the values of $\Sigma xy, \Sigma x^2y$, etc. The equations that would ordinarily result have been reduced to the form for the coefficients $a, b, c, \ldots$ of a series of equations of the form

$y = a + bx + cx^2 \ldots$

From the solutions given below, the coefficients a, b, c, .... may be obtained immediately by substituting the values of $\sum y$, $\sum xy$, etc. The final numerical coefficients are yielded by transforming the arbitrary $x = 1, 2, 3, \ldots$, or $(x-5) = -5, -4, -3, \ldots$, into the given values of x.

For the benefit of less experienced computers it may be pointed out that it is a little easier to multiply y times x, than yx times x, $yx^2$ times x, .... than y times $x^2$, y times $x^3$, etc.

A simple example will illustrate the procedure recommended and the closeness of its results to those of the direct procedure in general use. For a given set of ten observations, recorded in Table I., let $y = a + bx$, and say that the observed values of y correspond to $x = 0.5, 1.5, 2.5, \ldots 8.5, 9.5$.

In general, the Method of Least Squares, applied to a linear relationship, yields the following:

<u>Formulae for Calculating the Coefficients of</u> $y = a + bx$, Based on n <u>Observations</u>:

$$\left.\begin{aligned} a &= \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - (\sum x)^2} \\[2em] b &= \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \end{aligned}\right\} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

For our given observations these formulae lead to the equation

$$y = 3.0068181 + 1.9936364x \quad \ldots\ldots\ldots\ldots(A)$$

In Table I. the first two columns record the observed data; the third gives the values of y calculated by equation (A); the fourth gives $\Delta_1$, the differences between the calculated and the observed values of y,

From the solutions given below, the coefficients a, b, c, ... may be obtained directly by substituting the values of $\Sigma y$, $\Sigma xy$, etc. The exact numerical coefficients are yielded by transforming the arbitrary $x = 1, 2, 3, \ldots$, to $(x - k) = \ldots$, into the given values of x.

For the benefit of less experienced computers it may be noted that it is a little easier to multiply y times x, than yx times x, $yx^2$ times x, ... than y times $x^2$, y times $x^3$, etc.

A simple example will illustrate the procedure recommended and the closeness of the results to those of the direct procedure in general use. From a given set of ten observations, recorded in Table 1, let $y = a + bx$, and say that the observed values of y correspond to $x = 0.0, 1.0, 2.0, \ldots 9.0$.

In general, the method of least squares, applied to a linear relationship, yields the following:

<u>Formulae for calculating the coefficients of $y = a + bx$, based on n observations.</u>

$$a = \frac{\Sigma x^2 \Sigma y - \Sigma x \Sigma xy}{n \Sigma x^2 - (\Sigma x)^2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

$$b = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2}$$

For the given observations these formulae lead to the equation

$$y = 0.000000 + 1.0000018x \qquad \ldots\ldots\ldots\ldots(4)$$

In Table 1, the first two columns record the observed data; the third gives the values of y calculated by equation (4); the fourth gives the differences between the calculated and the observed values of y.

which yield the minimum: $\sum\Delta_1^2 = 0.168$. An additional fifth column shows the differences percent.

Table I.

| x | y(obs.) | y(calc.) | $\Delta_1$ (calc.-obs.) | $\Delta_1$ % |
|---|---|---|---|---|
| 0.5 | 4.10 | 4.0036363 | -0.0963637 | -2.35 |
| 1.5 | 6.15 | 5.9972727 | -0.1527273 | -2.48 |
| 2.5 | 7.80 | 7.9909091 | +0.1909091 | +2.45 |
| 3.5 | 9.85 | 9.9845455 | +0.1345455 | +1.37 |
| 4.5 | 12.10 | 11.9781819 | -0.1218181 | -1.01 |
| 5.5 | 13.80 | 13.9718183 | +0.1718183 | +1.25 |
| 6.5 | 15.90 | 15.9654547 | +0.0654547 | +0.41 |
| 7.5 | 18.05 | 17.9590911 | -0.0909089 | -0.50 |
| 8.5 | 20.10 | 19.9527275 | -0.1472725 | -0.73 |
| 9.5 | 21.90 | 21.9463639 | +0.0463639 | +0.21 |

We now employ the indirect procedure. The observations are plotted on a scale where 50 mm. represent one unit of x, and 20mm. one unit of y. Neighboring points are connected by straight lines. The arbitrary values x = 1, 2, 3, .... 8, 9, and the corresponding values of y read from the broken line are given in the first two columns of Table II. From these we get $\sum y = 116.90$ and $\sum xy = 704.35$, which yield immediately the coefficients a, b, of the required equation by substitution in the following formulae:

<u>Formulae for Calculating the Coefficients of</u> $y = a + bx$, <u>Based on Nine Points</u>:

x = 1, 2, 3, .... 8, 9.

$$\left.\begin{aligned} a &= \frac{+95\Sigma y - 15\Sigma xy}{180} \\ \\ b &= \frac{-15\Sigma y \quad 3\Sigma xy}{180} \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

We thus obtain the equation:

$$y = 3.0013889 + 1.9975000x \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (B)$$

Table II

| x | y | x(obs.) | y(obs.) | y(calc.) | $\Delta_2$ (calc.--obs.) | $\Delta_2$ % |
|---|---|---|---|---|---|---|
| | | 0.5 | 4.10 | 4.0001389 | -0.0998611 | -2.41 |
| 1 | 5.15 | | | | | |
| | | 1.5 | 6.15 | 5.9976389 | -0.1523611 | -2.48 |
| 2 | 7.00 | | | | | |
| | | 2.5 | 7.80 | 7.9951389 | +0.1951389 | +2.50 |
| 3 | 8.85 | | | | | |
| | | 3.5 | 9.85 | 9.9926382 | +0.1426389 | +1.45 |
| 4 | 11.00 | | | | | |
| | | 4.5 | 12.10 | 11.9901389 | -0.1098611 | -0.91 |
| 5 | 12.95 | | | | | |
| | | 5.5 | 13.80 | 13.9876389 | +0.1876389 | +1.36 |
| 6 | 14.85 | | | | | |
| | | 6.5 | 15.90 | 15.9851389 | +0.0851389 | +0.54 |
| 7 | 17.00 | | | | | |
| | | 7.5 | 18.05 | 17.9826389 | -0.0673611 | -0.37 |
| 8 | 19.10 | | | | | |
| | | 8.5 | 20.10 | 19.9801389 | -0.1198611 | -0.60 |
| 9 | 21.00 | | | | | |
| | | 9.5 | 21.90 | 21.9776389 | +0.0776389 | +0.35 |

The third and fourth columns of Table II, reproduce again for comparison the "observed" values of x and y; the fifth gives the values of y calulated by equation (B); the sixth shows $\Delta_2$, the differences between these calculations and the observations and, again, the last column shows the differences percent.

Plainly, the indirect procedure and equation (B) reproduce the results all but as well as the usual direct procedure and equation (A). The sum of the squares of the differences between the calculated and the observed values, $\sum\Delta_2^2 = 0.171$, is very close to the minimum, $\sum\Delta_1^2 = 0.168$, of the direct procedure.

In place of the nine-point formulae (1), the following based on eleven points, will be found more convenient in some cases:

Formulae for Calculating the Coefficients of $y = a + bx$, Based on Eleven Points: $x = 0, 1, 2, 3, \ldots 8, 9, 10$.

$$\left.\begin{aligned} a &= \frac{35\sum y - 5\sum xy}{110} \\ b &= \frac{-5\sum y + \sum xy}{110} \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

Applying these formulae to our test case, we obtain the equation:

$$y = 3.00454545 + 1.99636364x \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (C)$$

Table III. shows the results. The first two columns reproduce once more the "observed" values of x and y. The third gives the values of y calculated by equation (C). The fourth and fifth show the differences between the calculated and the observed values.

Here the sum of the squares of the differences, $\sum\Delta_3^2 = 0.169$, is even closer to the minumum $\sum\Delta_1^2 = 0.168$, yielded directly by the observations.

The differences between the values of y from our indirectly gotten equations (B) and (C) and those from equation (A) based immediately on the observations, are small compared with the errors of the observations themselves.

Table III.

| x | y(obs.) | y(calc.) | $\Delta_3$ (calc.--obs.) | $\Delta_3$ % |
|---|---|---|---|---|
| 0.5 | 4.10 | 4.00272727 | -0.09727273 | -2.37 |
| 1.5 | 6.15 | 5.99909091 | -0.15090909 | -2.45 |
| 2.5 | 7.80 | 7.99545455 | +0.19545455 | +2.51 |
| 3.5 | 9.85 | 9.99181818 | +0.14181818 | +1.44 |
| 4.5 | 12.10 | 11.98818182 | -0.11181818 | -0.92 |
| 5.5 | 13.80 | 13.98454545 | +0.18454545 | +1.34 |
| 6.5 | 15.90 | 15.98090909 | +0.08090909 | +0.51 |
| 7.5 | 18.05 | 17.97727273 | -0.07272727 | -0.40 |
| 8.5 | 20.10 | 19.97363636 | -0.12636364 | -0.63 |
| 9.5 | 21.90 | 21.97000000 | +0.07000000 | +0.32 |

Below are given several sets of formulae for calculating the coefficients of parabolic equations of the second, third, and fourth degrees, which will suffice to meet most ordinary needs. In using these formulae the number of significant figures in the products involved should only be reduced with great caution (if at all).

Formulae for Calculating the Coefficients of $y = a + bx + cx^2$, Based on Nine Points: $x = 0, 1, 2, 3, \ldots\ldots 7, 8.$

$$\left.\begin{aligned} a &= \frac{+3052\Sigma y - 1428\Sigma xy + 140\Sigma x^2y}{4620} \\ b &= \frac{-1428\Sigma y + 1037\Sigma xy - 120\Sigma x^2y}{4620} \\ c &= \frac{+140\Sigma y - 120\Sigma xy + 15\Sigma x^2y}{4620} \end{aligned}\right\} \ldots\ldots\ldots\ldots(4)$$

If the parabola is to pass through the origin, then

$$\left.\begin{aligned} a &= 0 \\ b &= \frac{+135\sum y - 19\sum xy}{660} \\ c &= \frac{-19\sum y + 3\sum xy}{660} \end{aligned}\right\} \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (7)$$

Formulae for Calculating the Coefficients of $y = a + bx + cx^2$, Based on Eleven Points: $x = 0, 1, 2, 3, \ldots\ldots 8, 9, 10.$

$$\left.\begin{aligned} a &= \frac{+4980\sum y - 1890\sum xy + 150\sum x^2y}{8580} \\ b &= \frac{-1890\sum y + 1078\sum xy - 100\sum x^2y}{8580} \\ c &= \frac{+150\sum y - 100\sum xy + 10\sum x^2y}{8580} \end{aligned}\right\} \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (8)$$

If the parabola is to pass through the origin, then

$$\left.\begin{aligned} a &= 0 \\ b &= \frac{+55\sum y - 7\sum xy}{330} \\ c &= \frac{-7\sum y + \sum xy}{330} \end{aligned}\right\} \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (9)$$

Formulae for Calculating the Coefficients of $y = A + B\,(x-5) + C(x-5)^2 + D(x-5)^3$, Based on Eleven Points: $(x-5) = -5, -4, -3, \ldots\ldots 3, 4, 5.$

$$\left.\begin{aligned} A &= \frac{+6408\sum y - 360\sum y(x-5)^2}{30888} \\ B &= \frac{+1865\sum y(x-5) + 89\sum y(x-5)^3}{30888} \\ C &= \frac{-360\sum y + 36\sum y(x-5)^2}{30888} \\ D &= \frac{-89\sum y(x-5) + 5\sum y(x-5)^3}{30888} \end{aligned}\right\} \quad \cdots\cdots\cdots\cdots\cdots\cdots \quad (10)$$

The coefficients of $y = a + bx + cx^2 + dx^3$ may be obtained directly by the following set of formulae; it will be noted that these still involve $(x-5) = -5, -4, -3, \ldots\ldots 4, 5$, and not $x = 0, 1, 2, 3, \ldots\ldots 10$.

<u>Formulae for Calculating the Coefficients of</u> $y = a+bx+cx^2+dx^3$ <u>Based on Eleven Points:</u> $x = 0, 1, 2, 3, \ldots\ldots\ldots 7, 9, 10$

$$\left.\begin{aligned}
a &= \frac{-2592\sum y + 1800\sum y(x-5) + 540\sum y(x-5)^2 - 180\sum y(x-5)^3}{30888} \\
b &= \frac{+3600\sum y - 4810\sum y(x-5) - 360\sum y(x-5)^2 + 286\sum y(x-5)^3}{30888} \\
c &= \frac{-360\sum y + 1335\sum y(x-5) + 36\sum y(x-5)^2 - 75\sum y(x-5)^3}{30888} \\
d &= \frac{-89\sum y(x-5) + 5\sum y(x-5)^3}{30888}
\end{aligned}\right\}\ldots(11)$$

As a rule, it will be simpler to use, not Formulae (11), but (10), then calculate the coefficients a, b, c, d, of $y = a+bx+cx^2+dx^3$ by: $a = A - 5B+25C-125D$; $b = B-10C+75D$; $c = C - 15D$; $d = D$.

If the cubic curve must pass through the origin, then:

$$\left.\begin{aligned}
a &= 0 \\
b &= \frac{-2310\sum y(x-5) + 390\sum y(x-5)^2 + 36\sum y(x-5)^3}{30888} \\
c &= \frac{+1085\sum y(x-5) - 39\sum y(x-5)^2 - 50\sum y(x-5)^3}{30888} \\
d &= \frac{-89\sum y(x-5) + 5\sum y(x-5)^3}{30888}
\end{aligned}\right\}\ldots\ldots\ldots\ldots(12)$$

Or else, still for the eleven points, $x = 0, 1, 2, 3, \ldots\ldots\ldots 9, 10$, the following

$$
\left.
\begin{aligned}
a &= 0 \\
b &= \frac{+36136\sum y - 11550\sum yx + 830\sum yx^2}{51480} \\
c &= \frac{-11550\sum y + 4125\sum yx - 315\sum yx^2}{51480} \\
d &= \frac{+830\sum y - 315\sum yx + 25\sum yx^2}{51480}
\end{aligned}
\right\} \ldots\ldots\ldots\ldots\ldots (13)
$$

<u>Formulae for Calculating the Coefficients of</u> $y = A+B(x-5)+C(x-5)^2+D(x-5)^3+E(x-5)^4$, <u>Based on Nine Points</u>: $(x-5) = 4, -3, -2, -1, 0, 1, 2, 3, 4$.

$$
\left.
\begin{aligned}
A &= \frac{+154656\sum y - 39960\sum y(x-5)^2 + 1944\sum y(x-5)^4}{370656} \\
B &= \frac{+42380\sum y(x-5) - 3068\sum y(x-5)^3}{370656} \\
C &= \frac{-39960\sum y + 18207\sum y(x-5)^2 - 1035\sum y(x-5)^4}{370656} \\
D &= \frac{-3068\sum y(x-5) + 260\sum y(x-5)^3}{370656} \\
E &= \frac{+1944\sum y - 1035\sum y(x-5)^2 + 63\sum y(x-5)^4}{370656}
\end{aligned}
\right\} \ldots\ldots (14)
$$

<u>Formulae for Calculating the Coefficients of</u> $y = A+B(x-5)+C(x-5)^2+D(x-5)^3+E(x-5)^4$, <u>Based on Eleven Points</u>: $(x-5) = -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5$.

$$\left.\begin{array}{l}
A = \dfrac{+41184\sum y - 6840\sum y(x-5)^2 + 216\sum y(x-5)^4}{123552} \\[2ex]
B = \dfrac{+7460\sum y(x-5) - 356\sum y(x-5)^3}{123552} \\[2ex]
C = \dfrac{-6840\sum y + 2019\sum y(x-5)^2 - 75\sum y(x-5)^4}{123552} \\[2ex]
D = \dfrac{-356\sum y(x-5) + 20\sum y(x-5)^3}{123552} \\[2ex]
E = \dfrac{+216\sum y - 75\sum y(x-5)^2 + 3\sum y(x-5)^4}{123552}
\end{array}\right\} \ldots\ldots\ldots\ldots\ldots (15)$$

<u>Formulae</u> <u>for</u> <u>Calculating</u> <u>the</u> <u>Coefficients</u> <u>of</u> $y = a+bx+cx^2+dx^3+ex^4$,

<u>Based</u> <u>on</u> <u>Eleven</u> <u>Points</u>: $x = 0, 1, 2, 3, \ldots\ldots 8, 9, 10.$

$$\left.\begin{array}{l}
a = \dfrac{+5184\sum y + 7200\sum y(x-5) - 3240\sum y(x-5)^2 - 720\sum y(x-5)^3 + 216\sum y(x-5)^4}{123552} \\[2ex]
b = \dfrac{-39600\sum y - 19240\sum y(x-5) + 17310\sum y(x-5)^2 + 1144\sum y(x-5)^3 - 750\sum y(x-5)^4}{123552} \\[2ex]
c = \dfrac{+25560\sum y + 5340\sum y(x-5) - 9231\sum y(x-5)^2 - 300\sum y(x-5)^3 + 375\sum y(x-5)^4}{123552} \\[2ex]
d = \dfrac{-4320\sum y - 356\sum y(x-5) + 1500\sum y(x-5)^2 + 20\sum y(x-5)^3 - 60\sum y(x-5)^4}{123552} \\[2ex]
e = \dfrac{+216\sum y - 75\sum y(x-5)^2 + 3\sum y(x-5)^4}{123552}
\end{array}\right\} 16$$

If this curve must pass through the origin, then:

$a = 0$

$$\left.\begin{array}{l}
b = \dfrac{-21600\sum y + 5760\sum y(x-5) + 6060\sum y(x-5)^2 - 1356\sum y(x-5)^3}{123552} \\[2ex]
c = \dfrac{+16560\sum y - 7160\sum y(x-5) - 3606\sum y(x-5)^2 + 950\sum y(x-5)^3}{123552} \\[2ex]
d = \dfrac{-2880\sum y + 1644\sum y(x-5) + 600\sum y(x-5)^2 - 180\sum y(x-5)^3}{123552} \\[2ex]
e = \dfrac{+144\sum y - 100\sum y(x-5) - 30\sum y(x-5)^2 + 10\sum y(x-5)^3}{123552}
\end{array}\right\} \ldots\ldots (17)$$

The saving of labor effected by our procodure will of course grow rapidly with the dogree of the equation desired.